学科领域研究前沿识别方法研究进展 *

■ 张雪[1,2] 张志强[1,2] 曹玲静[1,2] 阮伟南[1,2] 任晓亚[1,2] 冯志刚[1,2]
[1] 中国科学院成都文献情报中心 成都 610041 [2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘 要：[目的/意义]梳理国内外研究前沿相关成果，归纳总结现有研究存在的问题，为学科领域研究前沿识别提供参考借鉴。[方法/过程]首先对研究前沿识别的必要性进行归纳总结，其次对相关概念进行辨析，再次在调研国内外相关研究基础上从研究前沿识别方法研究、研究前沿识别新方向两个层面对其进行归纳整理，最后指出现有研究不足并对未来发展提出展望。[结果/结论]就概念界定而言，通过从时间维度和定义范围两方面辨析与研究前沿相关的系列概念，最终明确研究前沿的内涵。就识别方法而言，经典的研究方法包括直接引用、共被引分析、文献耦合以及基于词簇的研究前沿识别方法；同时，基于多源数据、多维指标以及机器学习算法的研究前沿识别是未来研究的新方向。在以上分析基础上，总结不同类型研究前沿识别方法的不足以及存在的普适性问题，并对未来研究重点进行展望。

关键词：研究前沿 专家判读 引文分析 词簇分析 多源数据 多维指标
分类号：G253
DOI：10.13266/j.issn.0252-3116.2022.12.013

1 引言

科学研究的本质是探索、求真[1]，科学的发展是一个进化与革命、积累与飞跃、连续和中断的不断循环往复的过程[2]，科学家要进行科技创新，首先需要明确本学科领域的焦点、痛点，但对科研人员来说，难点是如何确定、选择具有突破潜力的研究方向并开展实施。同时，根据《中国科技人才发展报告(2018)》和《中国科技人力资源发展研究报告(2018)》可知，中国的科技人力资源总量和研究与试验发展(R&D)人员全时当量持续增长，并连续 5 年位居世界首位，2017 年 R&D 人员总量达到 621.4 万人，2018 年底科技人力资源已达 10 154.5 万人。为了在“科学蛋糕”(资源、利益等)的分配中争得优先权，全世界科学家之间形成了愈加激烈的学术竞争。而科学前沿代表了未来科学发展的关键走向，从本质来看，科学前沿即为科学家手中的一张王牌，是提高科学家竞争优势的保障。因而，对研究前沿开展持续性的跟踪、监测与识别，尽早发现、预判出科学领域的前沿方向，能够为科学家把握研究焦点奠定基础，对未来顺利开展研究具有重要意义；有利于科研管理者从研究领域的本身价值层面(而非行政占优)配置资源，推动科学发展的良性生态；有利于决策者把握科学发展规律、动态调整政策机制导向，进而抢得科技发展先机。

鉴于研究前沿识别的现实意义，国内外学者也从不同角度出发对研究前沿进行了解读，但尚存在以定性方法为主、概念内涵未统一、指标体系多而杂等问题。本研究在调研国内外相关研究成果中发现，对于“研究前沿”和“新兴研究主题”两个概念，国内外学者经常将其等同或近似对待，因此在借鉴徐硕[3]、卢超[4]、张丽华[5]等的检索策略基础上，本研究以 Web of Science 数据库和中国知网为数据源，分别以 TS = ("research * front * " OR " scien * front * " OR " academic * front * " OR " frontier" OR " emerg * topic * " OR " emerg * research * topic * " OR " emerg * scien * topic * " OR " emerg * academic * topic * " OR " emerg * field * " OR " emerg * area * " OR " emerg * research * field * " OR " emerg * research * area * " OR " emerg * scien * field * " OR " emerg * scien * area * " OR " emerg * trend * ")和篇名 =(研究前沿 OR 新兴研究领域 OR 新兴

* 本文系国家社会科学基金重点项目“面向领域知识发现的学科信息学理论与应用研究”(项目编号:17ATQ008)研究成果之一。
作者简介：张雪，博士研究生；张志强，研究员，博士生导师，通信作者，E-mail：zhangzq@clas.ac.cn；曹玲静，博士研究生；阮伟男，博士研究生；任晓亚，博士研究生；冯志刚，博士研究生。
收稿日期：2021-12-06 修回日期：2022-02-21 本文起止页码：139-151 本文责任编辑：杜杏叶

研究趋势 OR 新兴研究主题 OR 新兴研究话题)为检索式,检索相关主题论文。进一步地,根据二八定律,结合被引频次,筛选国内外研究前沿识别重点文献,同时对参考文献中出现的有价值文献进行回溯检索,通过对上述重点文献的梳理和总结,首先对研究前沿相关概念进行辨析,讨论不同概念的区别和联系;其次从研究前沿识别方法、研究前沿识别新方向 2 个层面归纳已有研究内容;最后总结现有研究不足并对研究前沿识别未来发展做出展望。与已有研究前沿相关综述相比,本研究的创新之处在于:首先,在梳理研究前沿经典系列概念基础上,从时间维度和定义范围两方面对相关概念进行辨析介绍,进一步明晰各个概念之间异同,确定本研究所使用的概念体系;其次,除了梳理直接引用、共被引分析、文献耦合以及基于词簇的研究前沿识别常用方法外,补充了方法性能比较、方法间的融合和改进方面的相关研究;再次,对研究前沿识别的新方向进行总结。具体为:①对研究前沿识别中的数据源以及不同数据对象的特色进行归纳;②对研究前沿表征维度以及每一维度的测度指标进行梳理;③对文本分类、文本聚类、时间序列分析等机器学习算法在研究前沿识别中的新应用进行罗列;④从 5 个角度更加全面地对未来研究重点和思路进行展望。

2 研究前沿相关概念辨析

D. J. Price[6] 被认为是研究前沿(research front)领域的鼻祖,其于 1965 年首次提出研究前沿概念,他认为引文网络中最近的、被广泛引用的文献集合就是活跃的研究前沿,并将其形象地描述为“生长尖端”(growing tip)或“表皮层”(epidermal layer);1973 年,H. Small[7] 提出研究前沿是高被引论文聚类的结果,并将共被引聚类的方法用以识别研究前沿。至此,二者共同奠定了研究前沿领域的理论和方法基础,此后不同学者也从不同视角出发对研究前沿的概念进行补充和完善,代表性的定义如下:1991 年,R. R. Braam 等[8] 认为研究前沿是学者高密度关注的一系列主题,并利用共被引聚类间的相似性来探测研究前沿的稳定性;1994 年,O. Persson[9] 将被引文献视为“研究基础”,而将引用相同论文的施引文献簇定义为“研究前沿”;同年,E. Garfield[10] 将所有共被引聚类中的核心文献及其施引文献共同称为研究前沿;1998 年,S. Bhattacharya 等[11] 借助共词分析方法,直接从论文标题中抽取主题词进行共词聚类,将共词聚类所形成的研究主题视为研究前沿;2003 年,S. A. Morris 等[12] 将研究前沿与范式理论相结合,认为研究前沿是特定范式中引用一组持续性、相对固定的文献的论文集,强调被引文献的稳定性;2006 年,陈超美[13] 认为研究前沿可能是不连续的,是某个领域的暂时性问题,故将研究前沿定义为一组突显的动态概念和潜在的研究问题;2008 年,N. Shibata[14] 将最新的直接引文聚类定义为研究前沿;2010 年,S. P. Upham 等[15] 认为研究前沿是科学研究领域中最为动态变化和吸引科学家关注的研究主题,融合了科学发现和社会关注两个概念;2014 年,许晓阳等[16] 认为科学研究中最近出现、正在兴起的研究主题或研究领域就是研究前沿;2016 年,郑彦宁等[17] 认为研究前沿是特定研究领域和特定研究时间中最活跃的部分。整理研究前沿的相关概念见表 1。关于研究前沿的英文表述主要有"research front"和"research frontier",国内学者钟镇[18] 分别从理论和实证方面对两个术语进行了详细阐述:research front 是一个先验评价,是没有经过实际验证的期待结果,更多地出现在信息计量学;而 research frontier 是后验统计,是经过同行专家确认价值后的分析结果,更多地应用在自然科学。也有学者认为通过科技文献所识别出来的研究前沿更偏向于“研究焦点”或“研究热点”,而通常认为的研究前沿应当是类似于宇宙起源、生物演化、物质结构等少量尖端研究领域,其对应的英文是 research frontier。

表 1 “研究前沿”的代表性定义

时间	作者	定义
1965	D. J. Price	近期高被引文献的集合
1973	H. Small	共被引文献聚类
1991	R. R. Braam 等	施引文献聚类
1994	O. Persson	基于共被引的文献耦合聚类
1994	E. Garfield	被引文献和施引文献的集合
1998	S. Bhattacharya	共词聚类
2003	S. A. Morris 等	施引文献耦合聚类
2006	陈超美	突显的动态概念和潜在的研究问题
2008	N. Shibata	直接引文聚类
2010	S. P. Upham 等	小的高被引聚类
2014	许晓阳等	最近出现、正在兴起的研究主题或研究领域
2016	郑彦宁等	某段时间某个研究领域中最新出现、正在兴起并引起科学家高度关注的研究主题

综上,目前国内外学者分别从不同角度、基于不同原理对研究前沿进行了界定。对研究前沿的定义方式主要分为两大类:第一类是从科技文献数据角度界定研究前沿,分别有:被引文献、施引文献、文献本身三个层面,代表学者有 D. J. Price、H. Small、R. R. Braam、E.

Garfield、S. Bhattacharya、S. P. Upham 等；第二类则是从较为宏观的角度进行界定，不再受限于数据源，更加强调研究的动态性和活跃性，代表学者有陈超美、许晓阳、郑彦宁等。尽管学者们对研究前沿的界定方式存在差异，但研究前沿的内涵是清晰的：研究前沿是正在兴起的、具有发展潜力的、未来可能会引起大量关注的研究主题或研究领域，其具有新颖性、动态性、活跃性的特征。新颖性是指研究前沿应当是一个研究领域最先进的研究问题，代表了新兴的发展趋势；动态性是指研究前沿应当是随着时间变化而变化的，具有一定的时效性；活跃性是指研究前沿应是能够引起学术界高度关注的研究主题。

在信息科学领域，与研究前沿相似的概念有很多，如新兴研究、研究热点、科学前沿、学科前沿等。但研究前沿与这些相似概念在内涵上存在诸多差异，如与新兴研究、研究热点存在时间维度上的差异，与科学前沿、学科前沿存在定义范围上的差异。

（1）时间维度上的差异。"新兴研究"是指当下新出现的研究主题，主要特征在于"新"，突出时间新颖性。郭涵宁[19]将新兴研究解释为初次出现且蓬勃发展的研究，认为新兴研究强调的是当下，而研究前沿是在特定时间段内引起了广泛关注的新兴研究领域。罗瑞等[20]认为新兴研究虽然呈现"年轻化"和"快增长性"的趋势，但并不代表它在未来是具有研究价值和研究前景的研究前沿，即研究前沿应是有价值的、稳定的新兴研究。"研究热点"是指关注度比较高的研究主题，主要特征在于"热"，突出广泛讨论性。钟镇[18]指出，在时间轴上，前一时段具有学术价值的研究前沿将会有较大概率转化成新时段的研究热点，即研究热点相对于研究前沿在时间维度上具有一定的滞后性。综上，笔者认为研究前沿是指正在兴起的、被科学界高度关注的、研究内容具有一定创新性和发展潜力的研究主题，其在新颖性和关注性的基础上更加强调的是主题潜力，主要特征在于高创新性与高影响力。三者之间的关系如图 1 所示，在一定条件下，新兴研究会发展成为研究前沿，而部分研究前沿又会成为研究热点。具体来看，新兴研究是当下的研究探索，在未来可能会孕育出研究前沿；研究前沿在不断发展过程中，可能会引起学术共同体的广泛关注，进而发展成为研究热点。若新兴研究或研究前沿发展不利，则将"销声匿迹"，也就不能相应地发展成为研究前沿和研究热点。由此可见，研究热点具有明显的时间累积性和顺序性特征。

（2）定义范围上的差异。研究前沿与科学前沿、

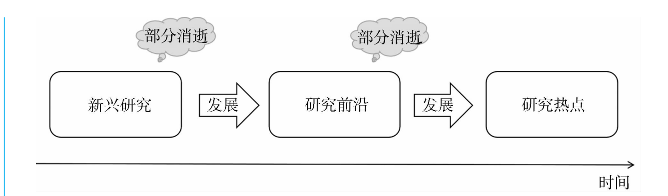


图 1　研究前沿与新兴研究、研究热点之间的关系

学科前沿的区别主要是在于研究领域和科学、学科之间的区别。"科学前沿"是指具有前瞻性、先导性、理论性、探索性，并对科学未来发展具有重大影响和引领作用的研究，又称为"科技前沿"。科学前沿是一个广义的概念，涵盖了所有与科学技术相关的学科、领域，而研究前沿则通常限定在一个特定研究领域。"学科前沿"是指某一学科中最有价值的发展趋势，一般是制约该学科当前发展的重大关键性问题，其讨论范围是学科。刘海峰[21]认为"学科"与"研究领域"的重要区别在于是否具有渗透性，学科的边界通常不可渗透，知识具有稳定性和整合性，而研究领域的边界是可渗透的，知识相对开放和松散。故随着学科越来越细化的划分，通常多个学科对应于一个研究领域。具体来看，三者在定义范围上的关系如图 2 所示，科学前沿是最为宏观的研究前沿，前沿问题的解决可能将科学研究带入新的发展阶段，对国民经济和社会发展都有重要意义。在大科学时代，很多科学问题的解决已经不再局限于某一个学科。因此，研究前沿是指某一个研究领域的前沿，有可能涉及多个学科，甚至是跨学科的。



图 2　研究前沿与科学前沿、学科前沿之间的关系

3　研究前沿识别主要研究内容

本文主要关注如何识别学科领域研究前沿，即从方法论的角度出发探讨采用何种方法、手段从学科领域主题中进一步抽取具有高新颖性、高发展潜力以及高影响力的主题，并将其用于推进有前瞻价值的课题开发。在调研国内外相关研究的基础上从研究前沿识别方法、新方向两方面对其进行归纳总结，其中识别方法研究从定性、定量两个视角对研究前沿概念提出至

今使用最为广泛、研究最为成熟的经典方法及其改进方法进行介绍,这部分研究是研究前沿识别的基石及后续研究得以成功开展的基础;新方向是在机器学习及大规模文本处理技术迅速发展的背景加持下,对已有研究前沿识别方法的新拓展,是未来研究的新思路。

3.1　研究前沿识别方法研究

自 D. J. Price 将研究前沿概念引入科技领域,学者就其概念内涵、识别方法进行了多方面的摸索和探讨,主要方法包括定性判读及定量计算,其中定量方法主要从引文网络分析角度展开。本小节在概述相关定性研究基础上,主要对定量分析进行总结,首先对研究前沿识别的 3 大视角:直接引用、共被引分析、文献耦合思想的提出、方法流程、经典研究进行概括;其次对 3 大方法性能对比、方法改进等研究进行归纳;以上分析均基于文献层面,存在固有缺陷,而基于词簇的分析方法可作为引文分析方法的补充或替代方法,故最后对研究前沿识别方法的另一个分支基于词簇的识别进展进行总结,以期更加全面、系统地梳理研究前沿识别方法体系脉络。

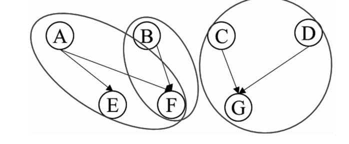
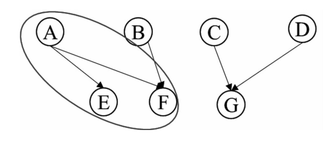
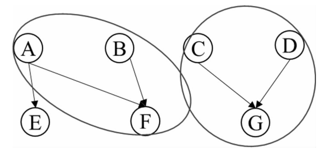
3.1.1　基于主观数据的专家判读研究前沿识别

德尔菲法广泛适用于专项及综合性科技研究的长期趋势预测,故目前德尔菲法是研究前沿识别领域最常用的方法之一。与德尔菲法相似的技术预见方法还包括头脑风暴法和专家咨询法,这两种方法不需要进行反复的问卷调查,可直接向专家征求意见,操作较为方便,大大节约了时间,但也容易造成识别结果缺乏民主化和社会化、准确性不能实现最大化等问题[22]。此外,也有大量研究将访谈与引文分析结果结合使用,并指出访谈是一种至关重要的意义构建工具[23-24]。如 S. Upham 等[15]共采访了 30 名研究人员,访谈时间为 30 分钟至 2 小时不等。目前,在研究前沿识别的研究中,无论前期核心文献集阈值的设定,还是后期量化分析结果的解读和修正均需专家的参与。但以上方法由于甄选专家的信息源和分析信息的方法存在缺陷,所以存在准确性及可靠性较低、客观性较差等缺陷。随着数据密集成为各学科的显著特点,如何使量化分析结果更好地辅助专家决策,专家在研究前沿识别中如何发挥作用、怎么发挥作用均是进一步研究的方向。

3.1.2　基于目标文献及其前向、后向引文的研究前沿识别

随着科技文献的迅速积累,科学计量方法成为定量识别研究前沿的重要手段。通过对已有研究前沿识别相关文献的调研发现,引文分析法是研究前沿识别中发展最早、理论基础最扎实、使用最广泛的方法之一。因引文分析按照不同的引用关系类型可划分为直接引用、共被引、文献耦合三种,故现有研究多从以上三种视角展开。表 2 为三种引文分析方法的含义辨析及其用于研究前沿识别的分析流程:

表 2　三种引文分析方法含义辨析及其用于研究前沿识别的分析流程

引文类型	直接引用(Direct Citation)	共被引分析(Co-Citation Analysis)	文献耦合(Bibliographic Coupling)
图解			
分析流程	①数据下载:确定研究领域和时间窗口,从数据库平台下载目标文献及其参考文献集合; ②引用对识别:识别数据集中的引用关系对,剔除没有引用或未被引用的文献; ③相似度计算:根据文献间引用频次大小计算相似度; ④聚类:根据相似度大小对目标文献进行聚类,并形成文献簇; ⑤命名:对不同文献簇进行主题命名	①数据下载:确定研究领域和时间窗口,从数据库平台下载目标文献及其施引文献集合; ②高被引目标文献筛选:因引用频次大小受时间影响,故对目标文献及其施引文献按年进行分组,对于每个年度集合,按照一定规则对目标文献进行筛选,如遴选满足引用时滞 = 0,被引频次 > = 3 或引用时滞 < 3,被引频次 > = 引用时滞 + 1,或被引频次 > = 5 的目标文献为高被引文献集合[25]; ③相似度计算:根据不同目标文献被施引文献共同引用频次计算两篇目标文献之间相似性; ④聚类:根据相似度大小对目标文献进行聚类,并形成共被引文献簇,若网络平均度中心度过高,可通过设置不同目标文献间相似度阈值大小对网络进行裁剪。 ⑤命名:对不同文献簇进行主题命名	①数据下载:确定研究领域和时间窗口,从数据库平台下载目标文献及其参考文献集合; ②数据筛选:剔除引用频次过高的参考文献,避免耦合过度聚合; ③相似度计算:根据不同目标文献同时引用相同参考文献频次计算两篇目标文献之间相似性; ④聚类:根据相似度大小对目标文献进行聚类,并形成文献耦合簇; ⑤命名:对不同文献簇进行主题命名

（1）直接引用（Direct Citation）。E. Garfield[26]于1963年指出直接引文分析法可作为评估科学发现影响力的关键方法；D. J. Price[1]以近期发表且频繁被直接引用的文献集为研究对象，对研究前沿主题进行了识别；R. Klavans等[27]指出直接引用可揭示某一领域的研究现状及未来发展趋势。但由于直接引用需要较长的时间窗口以获得足够的引用来保证聚类效果，故未能得到广泛使用。

（2）共被引分析（Co-Citation Analysis）。若论文E、论文F（不论其各自发表的时间）同时被论文A引用，则称论文E、论文F之间存在共被引关系，共被引强度是指同时引用论文E、论文F的论文篇数，篇数越多，则两篇共被引文献的相关程度越高。共被引是前瞻性的，共被引强度随时间推移可能发生变化。ESI一直采用共被引分析方法进行领域前沿预测，具体以成簇的、被频繁共同引用的高被引论文为研究对象。H. Small[7]于1973年提出可基于文献共被引关系探测研究前沿，他认为共被引关系比直接引用更能客观表征科学的智力和社会认知结构；1974年H. Small和B. C. Garfield等人[28－29]利用该方法对文献相似性进行了实证分析，并可视化展示了聚类结果；1985年，他对该方法进行了进一步修正，提出分数共被引聚类（Fractional Co-Citation Clustering）用以消除由于每篇论文的参考文献数目不同而对聚类结果的影响[30]；I. V. Marshakova[31]也于1973年指出相比于文献耦合（回溯性研究），共被引（前瞻性研究）分析方法更加复杂，也更能揭示研究前沿的演化特征。但基于共被引的研究前沿识别方法也存在一定的缺陷：研究前沿只有在施引文献数量达到一定规模时才可被监测出来，具有一定的时间滞后性，即共被引方法不能在某个研究前沿出现时实现立刻识别，而是只能在领域发展的某个后期发现它。共被引聚类是一种先验方法，以某领域高被引论文为研究对象，根据论文间共引模式进行聚类，但若某领域无论文被高度引用，则无法有效识别该领域的研究前沿。

（3）文献耦合（Bibliographic Coupling）。该词最先由Fano提出，M. M. Kessler[32]于1962年对相关概念进行了界定，他指出，文献耦合理论的基本出发点为共同引用一篇或多篇文献的两篇文献之间必存在相关关系，即若论文A、论文B同时引用了一篇或多篇相同的文章，则称论文A和论文B之间存在文献耦合关系。耦合强度是指引用相同文献的篇数，相同篇数越多，两篇耦合文献的相关程度越高。W. Glänzel等[33]认为通过耦合强度高的、刚发表论文的聚类可识别领域早期发展态势，比共被引分析更具优势，他通过文献耦合方法识别出了阿尔兹海默症、富勒烯等领域的研究前沿，经专家咨询证实该方法可为研究前沿的识别提供重要参考；M. H. Huang等[34]以2000－2009年的高被引论文为例，使用文献耦合方法探测了有机发光二极管的研究前沿，认为文献耦合方法识别出研究前沿与该领域专家的观点是契合的；S. A. Morris等[12]以炭疽数据为例，利用文献耦合方法识别并描绘了该领域研究前沿的演化趋势，技术预测专家小组认为该研究结果是有价值的。因文献耦合是回溯性的，耦合强度是固定不变的，故相比于共被引分析，该方法动态性较弱，且两篇文献可能引用同一文献的不同内容，故也可能造成耦合强度虚高的假象。

在3大方法相继被提出后，学者又从不同方法性能比较、方法的融合和改进方面进行了更深入的探讨，具体研究如下：

（1）对比分析以上3种研究方法性能。部分研究表明直接引用能得出更有意义的研究结果，如N. Shibata等[35]以氮化镓、复杂网络和碳纳米管为研究对象，对比分析了直接引用、共被引、文献耦合三种方法在识别领域研究前沿的优缺点，结果表明直接引用能得出更为准确的研究前沿，引文耦合比共被引更能监测出研究前沿。部分研究表明共被引分析识别效果更好，如J. Sharabchiev[36]以1981年免疫学主题相关文献为研究对象，对比分析了共被引和文献耦合方法所识别的研究前沿主题网络，结果显示共被引分析比文献耦合在绘制免疫学主题科学图谱方面表现更好。部分研究表明文献耦合识别效果更好，如B. Jarneving[37－38]认为相较于共被引分析，文献耦合得出的主题更加微观且更具有可解释性；M. H. Huang等[39]采用文献耦合和共被引两种方法分析有机发光二极管领域研究前沿的演变情况，结果表明两种方法均可用来追踪研究前沿的演变，但文献耦合能比共被引更早地、更多地识别研究前沿，性能更佳。部分研究则表明三种方法没有显著差异，如K. W. Boyack等[25]以生物医学领域文献为例，分别采用直接引用、共被引分析、文献耦合、引文－文本混合方法识别研究前沿，结果表明每种方法均可被认为是能够识别生物医学研究前沿的一种方法。总体来看，由于分析单元选择的不同、研究对象的差异等原因，以引文分析为基础的、不同的前沿识别方法可能得出不同的识别效果，但三种方法是可以互为补充的，以提供更为全面的领域研究前沿知识

结构。

(2)改进现有方法或方法间交叉融合。这部分的探讨旨在提高识别结果的可读性、精确度,主要切入点包括增加相关字段信息、划分不同时间窗口、融合多种研究方法等,具体研究包括:D. Zhao 等[40]提出作者耦合分析方法(Author-Bibliographic Coupling Analysis,ABCA),并将其应用于信息科学领域,通过将识别结果与作者早期提出的作者共被引分析方法[41](Co-Cited Authors,ACA)进行对比,结果显示二者各有所长,结合分析更有可能获得研究领域知识结构的全貌;C. Chen 等[42]提出了一套结合作者共引分析和参考文献共引分析的方法体系,可更灵活、更高效地命名共被引聚类类团;K. W. Boyack 等[43-44]基于共被引分析方法创建了高度详细、动态的全球科学地图,同时为了使研究前沿结果可解释性更强,其试图通过考量不同被引频次阈值、不同时间切片、不同布局算法、是否纳入文献耦合方法等多种情形以提高识别结果准确性。

由上文分析可知,共被引分析倾向于以发表年限较长的文章作为聚类对象,而不能有效涵盖尚未被引用的、发表年限较短的文章,文献耦合则倾向于以发表年限较短的文章作为聚类对象,而不能有效涵盖被引用的发表年限较长的文章。相比而言,直接引用则可在整个时间窗口内更均匀地对所有文献进行聚类。但限于人力、物力的影响,三种方法均以达到一定被引频次或耦合强度的文献为研究对象,不能全面分析可能与该研究前沿相关的所有文献。如 M. H. Huang 等[39]在研究中将耦合强度阈值设定为 5,ESI 按照总被引频次进行排序,提取排在每个 ESI 学科前 10% 的最具引文影响力的论文集作为研究对象。这可能会造成部分研究前沿相关文献的丢失,进而造成部分前沿主题的遗漏;其次,由于引用动机、引用位置等的不同,同一文献簇中的文献可能相似度较低,从而造成对前沿识别的误判;最后,三种方法均无法直接对类团进行命名,主题簇的命名方式大多还是基于对筛选出的主题文献题目、关键词以及摘要解读基础上的人为命名,存在较大的主观性,需要各领域专家学者对此进行进一步修正。

3.1.3 基于词簇的研究前沿识别

基于引文的研究前沿识别分析单元为高被引文献,低被引或零被引文献很难被纳入分析范畴。为克服引文分析的这一缺陷,部分学者开始将研究视角聚焦于更细粒度的词簇方向。鉴于学者对研究前沿的关注,某领域研究前沿出现后,将会随之涌现出大量的相关出版物,相关关键词的频次也会越来越高。学者试图利用词频统计的方法以更直接的方式发现更有价值的研究前沿。现有的相关识别方法主要包括基于词频(突发词)的前沿识别、基于共词的前沿识别,具体研究内容如下:

(1)基于词频(突发词)的前沿识别。随着突发词监测(Burst Term Detection,BTD)在文本挖掘中的大量应用,传统科学计量学开始使用突发词监测技术探究相关领域的研究前沿。突发词指文本流中频率突然激增的某个或某组单词,可用单词本身的频次变化和突发出现的时间间隔来表征[45]。J. Kleinberg[46]指出某主题出现是伴随着某些特征频率急剧上升的,即某领域研究主题的出现是存在"活动爆发"标志的,并开发突发词监测算法识别那些密度突然变大、词频突然变高的词;陈超美使用 Kleinberg 算法将突发词监测整合到 CiteSpace 中,并指出文献集中的突发词可部分展示某研究主题的潜在前沿[13];M. N. Li 等[47]为增强传统共词分析结果,引入突发词监测,构建了关键词与突发词间的关联规则挖掘模型,研究结果被证实可作为传统研究前沿识别的有效补充。与传统高频词不同,突发词更加强调频次突然增高的词,结合研究前沿的定义,传统高频词分析识别研究热点,突发词分析更可能识别出研究前沿。但突发词监测的关键在于能否准确识别出突发词汇,时间切片、频次阈值等的选择均会对监测效果产生重要影响,相同突发词监测算法应用到不同研究领域也可能得到不同的准确率,虽然研究者们一直在尝试开发不同算法以提高监测水平,但目前尚未形成一个普适性的监测算法。

(2)基于共词的前沿识别。因论文通常需要时间才能被其他论文所引用,故基于引文分析的研究前沿识别方法很难把握领域最新趋势,而共词网络在论文发表的第一时间即可迅速构建完成,故可及时发掘研究前沿。由于单个关键词可能会削弱研究主题的语义表达,故研究主题通常是由一组共现词簇构成,M. Callon 等[48]于 1986 年出版了第一部关于共词分析的学术专著,被认为是该研究领域的里程碑式工作。共词分析首先将数据集划分为不同时间段的子集,并以论文标题、摘要等内容的关键词作为研究对象;其次统计不同关键词对在论文中同时出现的频率;最后绘制各时间段的关键词对共现网络图谱。共词网络中的节点对应于关键词,边对应于关键词间的共现关系。共词分析方法通常用一般通用的、频繁出现的关键词对表征研究主题,如 J. Joung 等[49]通过层次聚类算法对

关键词相关矩阵进行聚类分析,结果表明可以通过关键词对识别研究前沿技术。但这些稳定的关键词对可能会干扰那些具有特异性的、突然引起注意的、具有特定时期特征的研究主题的发现,而这些爆炸性的研究主题往往更能代表领域的研究前沿。为克服这一弊端,M. Katsurai[50]开发了 TrendNets 算法,通过计算连续两个时间段关键词共现频率的差异以快速监测动态共词网络中边缘权值的变化,进而识别出那些在某一时间段被广泛讨论、而在之前时间段未被广泛讨论的主题。但目前研究中仍大多采用传统共词分析法,即利用关键词词频统计和高频词聚类识别某领域的研究前沿,该方法提取出的研究主题更偏向于研究热点。同时因共词分析同样基于关键词,忽略了文章的语义信息,故研究者开始尝试利用主题模型等机器学习方法提高研究前沿识别结果中的语义信息。

基于词簇的研究前沿识别方法所识别出的研究主题并不能被直接定义为研究前沿主题,需要专家的进一步判读或结合其他研究前沿识别方法。相关研究将词簇分析方法与引文分析方法相结合。首先通过引文分析方法识别出领域高被引文献集,其次以该文献集为研究对象,运用相关词簇分析方法挖掘出相关前沿主题。如 R. R. Braam 等[8]将词频分析方法与共被引分析方法相结合,通过对高被引文献集合标题、摘要的词频分析识别出研究前沿;侯海燕等[51]结合共词分析与共被引分析;周立英等[52]结合共词分析与引文耦合;P. Van den Besselaar 等[53]结合词频分析与共被引分析。不同方法间的结合能够有效弥补单一方法存在的缺陷,已成为目前研究前沿识别领域使用较多的方法。

3.2 研究前沿识别新方向

除上述常见的、主流的方法外,随着各类型数据库的不断完善、机器学习算法的兴起,学者在已有研究的基础上,对研究前沿识别方法进行多方面的探索,试图进一步丰富研究对象、提高主题有效性、可读性,进而使得识别结果更加贴合实际,更好服务决策。

3.2.1 基于多源数据的研究前沿研究对象

在科学研究进入第四范式的大背景下,基于大数据的知识发现成为科技领域知识发现的重要形式,这些数据一般以数据库的形式被规范存储以实现数据共享,最常见的数据库包括科技论文数据库、专利数据库,近年来一些科技项目数据库也逐渐落地。科学研究的研究主体为团队或个人,研究产出为团队或个人为解决实践问题集成多学科知识而产出的论文、专利、专著、项目等内容,故仅将论文作为主题识别的研究载体会使前沿识别结果存在局限性,也削减了其前瞻性价值。为改善这一研究局限,部分学者开始以融合的多源数据作为研究前沿识别的分析对象,虽然研究对象进一步丰富,但其研究前沿识别方法和思路基本一致,不同之处在于需根据数据的本质差别对具体指标数值、数值大小的意义以及识别出来的前沿类型进行区分。如白如江等[54]指出已发表的论文多为对未解的科学问题所做的种种探索并取得了重大或一定进步,这部分主题为"过去式"的研究前沿,而基金项目包含未解决的、近期正在进行的、具有明确研究目标和方法路径的研究前沿,为"进行时"的研究前沿,科技规划中则包括"未来时"的研究前沿;张婧等[55]以科技创新型国家重大科研项目数据资料归集为基础,从项目名称及关键词、项目摘要、项目所属科研计划三方面探索了基于科研项目数据的研究前沿;邓启平[56]以中国计算机协会推荐的人工智能领域 A 类、B 类期刊和会议论文为数据源,结合指标阈值遴选研究前沿;I. Park 等[57]以专利和论文为数据源,分别捕捉不同类型研究前沿;曾海娇等[58]基于专利与论文关联细粒度识别生物农药领域的潜在研究前沿。部分学者认为引文分析等方法不可避免地面临着时间滞后等缺陷,进而引入 Altmetrics 数据源以即时客观地反映论文影响力,结果表明以该方法识别出的主题与已有传统方法相比在社会关注层面具有更高的前沿性[59],归纳总结已有研究涉及的数据源见图 3。某领域制定的科技规划将会促进相关主题基金项目的申报,依托于基金项目也会有论文文献的产出,进而促进专利技术的产出、转化,但这一过程不是完全线性发展的,也可实现逆向促进,如基金项目或论文产出也会促进新的科技规划的制定。总体来看,目前研究前沿识别仍以论文数据为主要研究对象,未来可进一步扩大数据来源,比较同一领域不同数据源前沿主题识别的异同,进而识别出不同类型的研究前沿主题,提供更加有效的前沿方向。

3.2.2 基于多维指标的研究前沿衡量手段

一方面,有学者认为基于引文和词簇的研究前沿识别均侧重于"如何测度",而不是"如何识别"[3];另一方面,不同学者对于研究前沿的本质特征具有不同看法,有的学者认为,研究前沿应是在继承已有研究的基础上开创新的方法[60-61],有的学者则认为研究前沿应更多突出其对已有研究的破坏性[62-63],因此为了多维度全面识别研究前沿,学者试图根据研究前沿的概念内涵,设计不同维度的科学计量指标以分析不同前

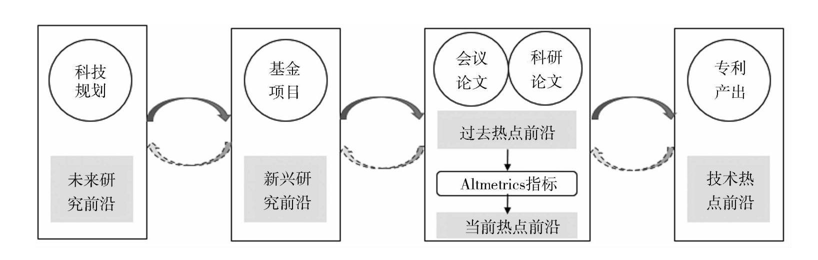


图3 研究前沿识别涉及的各类数据源

沿主题在各维度的表现，进而将前沿主题划分为不同类型，为决策支持提供更加对标、聚焦的政策建议。如S. Cozzens等[64]总结归纳了研究前沿的4个主要特征，分别为近期快速增长、对新事物有改变、市场或经济潜力较高以及科学性不断增长，该研究作为从多维指标视角测度研究前沿的开山之作，为研究前沿识别提供了新的思路。后续学者又从不同维度对该指标体系进行了补充和修改，如H. Guo等[65]认为特定单词频率突然增加、吸引新作者数量和频率加快、引用文献学科交叉性增强可作为一个领域是否为研究前沿的标志；H. Small等[61]认为学者对于研究前沿应具备新颖性、快速增长性几乎已达成共识；D. Rotolo等[66]根据研究前沿的概念内涵，清晰界定了研究前沿的5个特征，即高新颖性、相对快速增长、连贯性、显著影响、不确定性和模糊性；A. L. Porter团队[67-70]以新颖性、持续性、作者发文网络、增长性4个特征作为界定研究前沿的指标体系。国内学者在该方面也进行了系列研究，不过指标体系多为在上述研究基础上进一步的系统化。归纳梳理已有的研究前沿表征维度，并总结每一维度的测度指标，结果如表3所示。

通过对上述研究的进一步梳理，基于多维指标的研究前沿测度具体方法流程如下：首先采用上文或下文提到的研究前沿识别方法（如引文、词簇、文本聚类算法）识别出某领域研究前沿；其次根据研究前沿概念内涵构建测度框架，尽可能地使所提取的指标可以表达或全面衡量原数据的信息；最后根据指标重要性进行阈值划分，计算各主题在每个指标下的得分分值并将其划分为不同的前沿主题类型，如白如江等[54]结合主题强度和主题新颖度将研究主题划分为热点、新兴、衰弱、潜在四种不同的研究前沿主题；刘自强[71]结合主题新兴度和主题关注度，选择均位于前10%的主题为研究前沿主题；范云满等[72]利用发文量、被引量、新颖度曲线的交点表征主题发展的程度，将LDA主题模型识别结果与混合基线相比较，判读主题发展程度，进而识别出研究前沿主题。

综上，现有研究大多囊括了前沿主题识别的某个或某几个维度，具体以3-5个指标居多，复合指标计算复杂、难以推广，至今未能形成统一、系统的量化体系。相同维度测度指标的选择也有所不同，使得横向对比不同领域前沿识别结果几乎不可行。其次，因需人工指定各指标阈值范围，且各维度阈值设定的准确性也需进一步验证，这使得不同前沿主题类型的划分存在较多的主观性。

3.2.3 基于机器学习算法的研究前沿识别方法

机器学习的准确性和决策效率随着大规模文本处理技术的发展得以提高，近年来其使用率也呈指数级增长，机器学习算法应用于研究前沿识别领域主要包括文本聚类、文本分类、时间序列分析三方面，其中文本聚类主要采用非监督学习算法，并将其应用于文本主题识别，与传统共词分析相比，可降低关键词维度，提高主题词语义内涵和可解释性；文本分类主要采用监督学习算法，提前预测高被引文献集合，有效解决引文时滞问题；时间序列分析主要以测度指标随时间变化趋势为数据基础，通过时间序列模型预测指标未来发展趋势。具体研究如下：①基于文本聚类算法的前沿主题识别。随着数据体量的增大、非结构化数据的增多，文本主题挖掘技术的重要性日益突显。文本主题挖掘指从结构化、半结构化或非结构化的文本数据中获取有价值的信息和知识，主要包括文本收集、数据清洗、特征提取、特征修剪、文本聚类等流程，为了考虑词与词之间的上下文关系，学者在特征提取、文本聚类等方面开发不同的算法，如J. Yoon等[73]提出了SAO结构分析方法，与词簇分析方法相比，按语法结构组织的句子能清晰描述句子组成部分之间的关系[74]；接着不同学者也将其应用于研究前沿识别中，识别主题更易于理解[75-76]；W. M. Pottenger等[77]利用神经网络模型识别出了数据集中新出现的概念或主题；A. Kontostathis等[78]提出了Emerging Trend Detection方法，即首先通过共词分析展示不同时间段内的主题，其次使用文本挖掘技术根据共现特征进行主题抽取、分类，最

表 3　研究前沿识别多维测度指标

类型	指标	介绍
新颖性测度	即时性	探析某主题参考文献及其引用的知识基础被提出或正式发表时间
	新颖度	若某主题内文献集群发表年份越新，说明该研究主题新颖程度越高
增长力测度	发文量	以某主题内发文年增长率来表征
	引用影响力	以某主题内发文被引增长率来表征
	作者数量	以某主题内发文作者增长率来表征
影响力测度	引用影响力	以某主题内发文被引频次来表征
	创新性	以某主题内新发表文献与已有文献的知识冗余度来表征
	作者数量	以某主题内年度发文作者数量来表征
公众认可测度	普遍认可性	某主题在社交媒体平台被关注、点赞、转发次数；或在学术论文、报告等被提及、引用次数
	权威性	通过主题被提及的社交媒体或社交用户的影响力、刊载学术论文或报告的期刊或科研人员影响力来表征
交叉程度测度	学科多样性	以学科丰富性、学科平衡性、学科差异性来表征
	网络凝聚性	以某主题所涉及的学科网络紧密程度和各学科在网络中位置的差异程度来表征，具体指标包括包括网络密度、网络中心势、核心－边缘度等
其他二级指标	生命周期特征	以某主题下不同主题词在时间轴上的演化趋势来表征
	基本科学指标(ESI)	通过对近十年内高校及科研机构的论文统计分析，遴选热点和高被引论文，以此为数据基础，剖析领域研究前沿
	自然指数(Nature Index,NI)	由自然(Springer Nature)旗下的自然科研(Nature Research)出版，通过追踪高质量自然科学期刊所发表的科研论文作者信息，为科研共同体提供关于世界范围内科学研究现状及出版动向的信息
	π 指数(Productivity Index)	综合同行评议和论文影响力等定性、定量指标

后采用一定评价标准验证主题并判断其发展趋势；为了丰富主题的语义结构，D. M. Blei 等[79]提出使用 LDA 主题模型以无监督方式对数据集中隐含的语义结构进行挖掘；T. Mikolov 等[80]提出词嵌入模型 Word2vec 学习词的隐含向量表示；S. Xu 等[81]使用主题 n-grams 模型提取基于术语的主题。以上衍生的多种不同算法均使得研究前沿主题识别结果语义更加丰富，有助于专家的进一步理解和解读。②基于文本分类的前沿数据集预测。我们可以看到，前述研究大多基于高被引论文，但论文被引量的累积需要一定时间，故高被引论文集中不能有效涵盖最新发表但未来可能会被大量引用的文献，进而导致识别出的研究前沿新颖性无法保证，因此采用机器学习分类模型及早预测价值高、潜力强的文献为当下及未来研究前沿识别提供新的研究方法。如 C. Lee 等[82]基于专利数据构建 18 个高价值专利判别指标，采用前馈多层神经网络模型捕捉输入和输出指标间关系，进而在专利申请早期阶段提前预判具有研究前沿特性的专利；李欣等[83]首先通过构建机器学习模型来识别潜在高被引论文，其次以高被引论文集为数据源，利用聚类分析法识别研究前沿主题。③基于时间序列分析的指标演化预测。目前该方面的研究处于初步阶段，主要研究如 S. Xu 等[84]将研究前沿指标界定为新颖性、一致性等方面，分别计算各主题不同指标 2001－2017 年演化趋势，采用时间序列分析模型多任务最小二乘支持向量机预测指标未来 2 年变化趋势，进而预测潜在研究前沿主题；与上述研究思路相似，李静等[85]、岳丽欣等[86]分别采用支持向量机、ARIMA 模型对主题趋势预测。

从基于词频统计发展至文本聚类算法，从基于主观设定一定阈值的高被引文献集发展至提前预测潜在高被引文献集，从基于特定时间段指标计算发展至指标未来演化趋势预测，新的研究前沿识别方法使得研究主题语义信息更加丰富、识别粒度更加灵活，同时克服了引文分析的时间滞后性缺点，而且识别出来的主题更加具有前瞻性，为研究前沿识别提供了新的研究思路和测度方法。

4　总结与展望

本文从研究前沿发展背景、相关概念辨析入手，以研究前沿识别方法、研究前沿识别关键技术为视角，梳理总结了从定性、定量方法识别研究前沿主题的相关内容及目前研究前沿识别中使用的测度指标、研究对象，综合现有实践和研究，提出以下不足并对未来发展提出展望：

4.1　研究前沿相关概念需进一步明晰

目前研究前沿的概念尚不明晰，更多是从文献的角度界定和发现研究前沿，但这样识别出来的“研究前沿”是否是真正意义上的研究前沿还有待进一步探讨，有学者认为目前通过文献数据所测度识别出的研究前沿更像是研究焦点。但毋庸置疑的是，研究前沿的识

别对预测科学发展趋势及学者科研方向选择都有重要意义。故本文对研究前沿的相关概念进行了辨析，试图厘清研究前沿、研究热点、新兴研究、科学前沿及学科前沿之间的关系，为相关学者开展研究前沿工作提供参考。

4.2 阈值设置的合理性、有效性需进一步验证

研究前沿识别一般首先通过设定被引频次阈值遴选具有代表性的高被引文献为研究对象，低被引或零被引文献很难被纳入分析，如 ESI 研究前沿报告将在同出版年、同学科论文中位居前 1% 的论文作为高被引核心文献集，Upham 等[15]以各学科被引频次排名前 1% 的论文为研究对象，该参数可根据分析所需向上或向下调整，Glänzel[33]将文献链接超过 9 个，链接强度至少为 0.25 的文献定义为核心文献，Huang 等[34]将耦合强度阈值设置为 5。由于存在时间滞后，一篇文章达到高引用率需要多年时间，且不同学科情况不同；其次阈值不同，前沿主题识别结果也会存在差异。如何确定阈值大小一直没有明确的科学依据，一般由研究者根据数据体量的大小、前沿主题数目的个数来主观确定，故阈值设置的科学性、合理性都需进一步探讨和完善。

4.3 研究方法的适用范围需进一步界定

如果我们把研究前沿的本质看作是面向未来的探索，那么研究前沿应该是一个动态的、多元的、多维的概念。但目前研究前沿识别数据来源多为论文及其引文数据，同时为确定研究前沿演化趋势，研究者通常会设置引文窗口，将研究时间范围划分为若干区间，由于没有准确的、公认的引文窗口，研究者通常根据自己的研究目的选择引文窗口，5 年固定窗口最常用于研究前沿的研究。一方面被引次数受发表时间、作者引用动机、文章可获得性等因素的影响，难以逾越时滞性问题，故现有研究识别出的主题是否可以称之为真正意义上的研究前沿主题仍需商榷；另一方面由于突发词、睡美人文献的存在，固定时间窗口的划分能否有效囊括某领域全部研究前沿均需进一步思考。未来研究中可考虑融合多源数据，捕捉同一领域不同数据源前沿识别主题的相同或互补特征，同时可考虑在滑动窗口下的主题演化趋势中识别该领域前沿主题。

4.4 多源融合数据、多维测度指标需进一步系统化

现有研究主要以论文数据为研究对象，部分研究加入了基金项目、科技规划、专利等补充数据源。不同数据源具有不同的数据特色，有研究表明同一领域论文的研究主题比基金项目研究主题平均滞后 2 年，但现有研究主要将不同数据源进行简单的组合，分别挖掘各数据源的研究主题，然后结合数据源自身的优势将前沿主题划分为不同类型，忽略了由于数据源本身的差异所导致的前沿主题的不同。故如何纳入不同数据源的差异与特色，将同一领域不同数据源有效融合，基于融合后的数据矩阵进行聚类是弥补识别结果时滞性问题的突破口之一。多维测度指标方面，全面考虑研究前沿概念内涵本质的多维指标较少，大多指标计算复杂、难以推广，且至今尚未形成统一、系统的量化体系，故应进一步明晰研究前沿本质，结合不同特征，在指标尽可能反映研究前沿全要素基础上简化计算方法，使定量分析结果最大化地客观反映研究前沿主题的本质。

4.5 现有研究前沿识别结果的针对性、价值性需进一步明确

首先，研究前沿最常见的共同特征是它有可能改变并为我们对某一问题的认知注入新的理解，从某种程度来说，研究前沿并不是完全可计量的，也就是说，使用科学计量方法识别出的领域研究前沿主题不可避免地存在各种问题；其次，目前使用的研究前沿识别方法不论研究前期、后期均会涉及专家的参与和打分，但某领域科学共同体对于该领域研究前沿本身就有自己的认知和理解，因此，从科学计量角度识别出的研究前沿对于这些研究人员而言真正的意义和价值体现在哪里需要我们进一步明确。综上，我们应该明确服务对象和服务目的，科学计量方法主要是从宏观层面提供相对更加客观的趋势性信息，需要领域专家以计量结果为基础，进一步判读计量数据背后隐含的信息。同时对于识别的研究前沿主题范畴来说，我们应聚焦于识别研究者可能会忽略的、但具有潜在价值的研究前沿，希望我们的研究可以抛砖引玉，吸引更多研究者更加关注某些主题，也为研究课题的选择、政策的制定提供参考。

参考文献：

[1] ARTS S, VEUGELERS R. The technological origins and novelty of breakthrough inventions[R/OL]. FEB Research Report-MSI_1302. https://lirias.kuleuven.be/1830744? limo=0, 2013.
[2] 库恩. 科学革命的结构[M]. 金吾伦,胡新和,译.北京:北京大学出版社, 2003.
[3] XU S, HAO L, AN X, et al. Review on emerging research topics with key-route main path analysis[J]. Scientometrics, 2020, 122(1): 607 -624.
[4] 卢超, 侯海燕, DING Y,等. 国外新兴研究话题发现研究综述[J]. 情报学报, 2019, 38(1):97 -110.
[5] 张丽华. 研究前沿探测及其演化分析方法与实证研究[D]. 北京: 中国科学院大学, 2015.
[6] PRICE D J D. Networks of scientific papers[J]. Science, 1965,

149(3683):510－515.
[7] SMALL H. Co-citation in the scientific literature: a new measure of the relationship between two documents[J]. Journal of the American Society for Information Science, 1973, 24(4):265－269.
[8] BRAAM R R, MOED H F, VAN RAAN A F J. Mapping of science by combined co-citation and word analysis[J]. Journal of the American Society for Information Science, 1991, 42(4):233－251.
[9] PERSSON O. The Intellectual Base and Research Fronts of JASIS 1986－1990[J]. Journal of the American Society for Information Science, 1994, 45(1):31－38.
[10] GARFIELD E. Research fronts[J]. Current contents, 1994(41):3－7.
[11] BHATTACHARYA S, BASU P K. Mapping a research area at the micro level using co-word analysis[J]. Scientometrics, 1998, 43(3):359－372.
[12] MORRIS S A, YEN G, WU Z, et al. Time line visualization of research fronts[J]. Journal of the American Society for Information Science and Technology, 2003, 54(5): 413－422.
[13] CHEN C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3):359－377.
[14] SHIBATA N, KAJIKAWA Y, TAKEDA Y, et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications[J]. Technovation, 2008, 28(11):758－775.
[15] UPHAM S, SMALL H. Emerging research fronts in science and technology: patterns of new knowledge development[J]. Scientometrics, 2010, 83(1):15－38.
[16] 许晓阳，郑彦宁，赵筱媛，等. 研究前沿识别方法的研究进展[J]. 情报理论与实践，2014，37(6):139－144.
[17] 郑彦宁，许晓阳，刘志辉. 基于关键词共现的研究前沿识别方法研究[J]. 图书情报工作，2016，60(4): 85－92.
[18] 钟镇. 从高被引与零被引论文的引文结构差异看 Research Front 与 Research Frontier 的区别[J]. 图书情报工作，2015，59(8):87－96.
[19] 郭涵宁. 多元科学指标视角下的新兴研究领域识别探索[D]. 大连:大连理工大学，2013.
[20] 罗瑞，许海云，董坤. 领域前沿识别方法综述[J]. 图书情报工作，2018，23(62):119－131.
[21] 刘海峰. 高等教育学：在学科与领域之间[J]. 高等教育研究，2009，30(11):45－50.
[22] 沙振江，张蓉，刘桂锋. 国内技术预见方法研究述评[J]. 情报理论与实践，2015，38(6):140－144,120.
[23] CUHLS K. From forecasting to foresight processes-new participative foresight activities in Germany[J]. Journal of forecasting, 2003, 22(2 - 3):93－111.
[24] SMALL H. A co-citation model of a scientific specialty: a longitudinal study of collagen research[J]. Social studies of science, 1977, 7(2):139－166.
[25] BOYACK K W, KLAVANS R. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately? [J]. Journal of the American Society for Information Science and Technology, 2010, 61(12): 2389－2404.
[26] GARFIELD E. Citation indexes in sociological and historical research[J]. American documentation, 1963, 14(4):289－291.
[27] KLAVANS R, BOYACK K W. Identifying a better measure of relatedness for mapping science[J]. Journal of the American Society for Information Science and Technology, 2006, 57(2):251－263.
[28] SMALL H, GRIGGITH B C. The structure of scientific literatures I: identifying and graphing specialties[J]. Science studies, 1974, 4(1):17－40.
[29] GRIFFITH B C, SMALL H G, STONEHILL J A, et al. The structure of scientific literatures II: toward a macro-and microstructure for science[J]. Science studies, 1974, 4(4):339－365.
[30] SMALL H, SWEENEY E, GREENLEE E. Clustering the Science Citation Index using co-citations[J]. Scientometrics, 1985, 8(5/6):321－340.
[31] MARSHAKOVA I V. System of document connections based on references[J]. Nauchno-tekhnicheskaya informatsiya seriya 2-informatsionnye protsessy I sistemy, 1973(6):3－8.
[32] KESSLER M M. Bibliographic coupling between scientific papers [J]. American documentation, 1963, 14(1):10－25.
[33] GLÄNZEL W, CZERWON H. A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level[J]. Scientometrics, 1996, 37(2):195－221.
[34] HUANG M H, CHANG C P. Detecting research fronts in OLED field using bibliographic coupling with sliding window[J]. Scientometrics, 2014, 98(3):1721－1744.
[35] SHIBATA N, KAJIKAWA Y, TAKEDA Y, et al. Comparative study on methods of detecting research fronts using different types of citation[J]. Journal of the American Society for Information Science and Technology, 2009, 60(3):571－580.
[36] SHARABCHIEV J. Cluster analysis of bibliographic references as a scientometric method[J]. Scientometrics, 1989, 15(1/2):127－137.
[37] JARNEVING B. A comparison of two bibliometric methods for mapping of the research front[J]. Scientometrics, 2005, 65(2): 245－263.
[38] JARNEVING B. Bibliographic coupling and its application to research-front and other core documents[J]. Journal of informetrics, 2007, 1(4):287－307.
[39] HUANG M H, CHANG C P. A comparative study on detecting research fronts in the organic light-emitting diode (OLED) field using bibliographic coupling and co-citation[J]. Scientometrics, 2015, 102(3):2041－2057.
[40] ZHAO D, STROTMANN A. Evolution of research activities and intellectual influences in information science 1996－2005: introducing author bibliographic-coupling analysis[J]. Journal of the A-

merican Society for Information Science and Technology, 2008, 59(13):2070 – 2086.

[41] ZHAO D, STROTMANN A. Information science during the first decade of the Web: an enriched author co-citation analysis[J]. Journal of the American Society for Information Science and Technology, 2008, 59(6):916 – 937.

[42] CHEN C, IBEKWE-SANJUAN F, HOU J. The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis[J]. Journal of the American Society for Information Science and Technology, 2010, 61(7):1386 – 1409.

[43] BOYACK K W, KLAVANS R, SMALL H, et al. Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science[J]. Journal of engineering and technology management, 2014, 32:147 – 159.

[44] BOYACK K W, KLAVANS R. Creation of a highly detailed, dynamic, global model and map of science[J]. Journal of the Association for Information Science and Technology, 2014, 65(4):670 – 685.

[45] LEE S, PARK Y, YOON W C. Burst analysis for automatic concept map creation with a single document[J]. Expert systems with applications, 2015, 42(22):8817 – 8829.

[46] KLEINBERG J. Bursty and hierarchical structure in streams[J]. Data mining and knowledge discovery, 2003, 7(4):373 – 397.

[47] LI M N, CHU Y Q. Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis[J]. Journal of information science, 2017, 43(6):725 – 741.

[48] CALLON M, COURTIAL J P, TURNER W A, et al. From translations to problematic networks: an introduction to co-word analysis[J]. Social science information, 1983, 22(2):191 – 235.

[49] JOUNG J, KIM K. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data[J]. Technological forecasting and social change, 2017, 114:281 – 292.

[50] KATSURAI M, ONO S. TrendNets: mapping emerging research trends from dynamic co-word networks via sparse representation[J]. Scientometrics, 2019, 121(3):1583 – 1598.

[51] 侯海燕，刘则渊，栾春娟. 基于知识图谱的国际科学计量学研究前沿计量分析[J]. 科研管理, 2009, 30(1):164 – 170.

[52] 周丽英，冷伏海，左文革. 引文耦合增强的共词分析方法改进研究——以 ESI 农业科学研究主题划分为例[J]. 情报理论与实践, 2015, 38(11):120 – 125.

[53] VAN DEN BESSELAAR P, HEIMERIKS G. Mapping research topics using word-reference co-occurrences: a method and an exploratory case study[J]. Scientometrics, 2006, 68(3):377 – 393.

[54] 白如江，刘博文，冷伏海. 基于多维指标的未来新兴科学研究前沿识别研究[J]. 情报学报, 2020, 39(7):747 – 760.

[55] 张婧，刘彦君，张炜，等. 基于科研项目数据的科技前沿识别有效路径实证探索[J]. 科技管理研究, 2019, 39(16):108 – 119.

[56] 邓启平，陈卫静，张玲玲，等. 基于多维特征测度的人工智能领域研究前沿分析[J]. 情报杂志, 2020, 39(3):56 – 62.

[57] PARK I, LEE K, YOON B. Exploring promising research frontiers based on knowledge maps in the solar cell technology field[J]. Sustainability, 2015, 7(10):13660 – 13689.

[58] 曾海娇，孙巍. 基于专利与论文关联的潜在科学前沿识别——以生物农药领域为例[J]. 农业展望, 2020, 16(9):93 – 100.

[59] 王菲菲，刘明. Altmetrics 视角下的交叉学科研究前沿探测——以医学信息学领域为例[J]. 情报学报, 2020, 39(10):1011 – 1020.

[60] KLAVANS R, BOYACK K W. Using global mapping to create more accurate document-level maps of research fields[J]. Journal of the American Society for Information Science and Technology, 2011, 62(1):1 – 18.

[61] SMALL H, BOYACK K W, KLAVANS R. Identifying emerging topics in science and technology[J]. Research policy, 2014, 43(8):1450 – 1467.

[62] AZOULAY, P. Small research teams 'disrupt' science more radically than large ones[J]. Nature, 2019, 566(7744):330 – 332.

[63] BORNMANN L, TEKLES A. Disruptive papers published in Scientometrics[J]. Scientometrics, 2019, 120(1):331 – 336.

[64] COZZENS S, GATCHAIR S, KANG J, et al. Emerging technologies: quantitative identification and measurement[J]. Technology analysis & strategic management, 2010, 22(3):361 – 376.

[65] GUO H, WEINGART S, BÖRNER K. Mixed-indicators model for identifying emerging research areas[J]. Scientometrics, 2011, 89(1):421 – 435.

[66] ROTOLO D, HICKS D, MARTIN B R. What is an emerging technology? [J]. Research policy, 2015, 44(10):1827 – 1843.

[67] GARNER J, CARLEY S, PORTER A L, et al. Technological emergence indicators using emergence scoring[C]//2017 Portland international conference on management of engineering and technology. IEEE, 2017: 1 – 12.

[68] CARLEY S F, NEWMAN N C, PORTER A L, et al. An indicator of technical emergence[J]. Scientometrics, 2018, 115(1):35 – 49.

[69] PORTER A L, GARNER J, CARLEY S F, et al. Emergence scoring to identify frontier R&D topics and key players[J]. Technological forecasting and social change, 2019, 146:628 – 643.

[70] WANG Z, PORTER A L, WANG X, et al. An approach to identify emergent topics of technological convergence: a case study for 3D printing[J]. Technological forecasting and social change, 2019, 146:723 – 732.

[71] 刘自强. 基于主题扩散演化滞后的研究前沿识别研究[D]. 北京:中国科学院大学, 2020.

[72] 范云满，马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究[J]. 情报学报, 2014, 33(7):698 – 711.

[73] YOON J, PARK H, KIM K. Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis[J]. Scientometrics, 2013, 94(1): 313 – 331.

[74] CASCINE G, ZINI M. Measuring patent similarity by comparing inventions functional trees[M]. Boston: Springer, 2008.

[75] 李欣，谢前前，黄鲁成，等. 基于 SAO 结构语义挖掘的新兴技术演化轨迹研究[J]. 科学学与科学技术管理，2018，39(1)：17 – 31.

[76] 黄鲁成，张璐，吴菲菲，等. 基于突现文献和 SAO 相似度的新兴主题识别研究[J]. 科学学研究，2016，34(6):814 – 821.

[77] POTTENGER W M，YANG T. Detecting emerging concepts in text data mining [M]//BERRY M. Computational information retrieval. Philadelphia：Society for Industrial and Applied Mathematics，2001:89 – 105.

[78] KONTOSTATHIS A，GALITSKY L M，POTTENGER W M，et al. A survey of emerging trend detection in textual data mining[A]//Survey of text mining[M]. New York：Springer，2004:185 – 224.

[79] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research，2003，3:993 – 1022.

[80] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[J]. ArXiv preprint arXiv：1301.3781，2013.

[81] XU S，HAO L，YANG G，et al. A topic models based framework for detecting and forecasting emerging technologies[J]. Technological forecasting and social change，2021，162:120366.

[82] LEE C，KWON O，KIM M，et al. Early identification of emerging technologies：a machine learning approach using multiple patent indicators[J]. Technological forecasting and social change，2018，127:291 – 303.

[83] 李欣，温阳，黄鲁成，等. 一种基于机器学习的研究前沿识别方法研究[J]. 科研管理，2021，42(1):20 – 32.

[84] XU S，HAO L，AN X，et al. Emerging research topics detection with multiple machine learning models[J]. Journal of informetrics，2019，13(4):100983.

[85] 李静，徐路路，赵素君. 基于时间序列分析和 SVM 模型的基金项目新兴主题趋势预测与可视化研究[J]. 情报理论与实践，2019，42(1):118 – 123,152.

[86] 岳丽欣，刘自强，胡正银. 面向趋势预测的热点主题演化分析方法研究[J]. 数据分析与知识发现，2020，4(6):22 – 34.

作者贡献说明：

张雪：进行文献调研、研究资料收集与分析，撰写与修订论文；
张志强：提出论文研究思路，参与论文修订；
曹玲静：进行文献调研、研究资料收集与分析，参与论文修订；
阮伟男：进行文献调研、研究资料收集；
任晓亚：进行文献调研、研究资料收集；
冯志刚：进行文献调研、研究资料收集。

Research Progress of Research Front Recognition Methods in Subject Fields

Zhang Xue[1,2] Zhang Zhiqiang[1,2] Cao Lingjing[1,2] Ruan Weinan[1,2] Ren Xiaoya[1,2] Feng Zhigang[1,2]

[1] Chengdu Library and Information Center，Chinese Academy of Sciences，Chengdu 610041

[2] Department of Library，Information and Archives Management，School of Economics and Management，University of Chinese Academy of Sciences，Beijing 100190

Abstract：[**Purpose/Significance**] Sorting out the relevant researches of domestic and foreign research front，this paper summarizes the problems existing in the existing researches，and provides references for the identification of the research front in the subject field. [**Method/Process**] This paper first summarized the necessity of research front identification，then discriminated the relevant concepts. Subsequently，based on the investigation of domestic and foreign relevant researches，this paper classified it from two aspects of the research methods of research front identification and the new direction of research front identification，and finally put forward the existing research deficiencies and prospects for the future development. [**Result/Conclusion**] In terms of concept definition，the connotation of the research front is finally clarified by analyzing the series of concepts related to the research front from two perspectives of the time dimension and the definition scope. In terms of identification methods，classical research methods include direct citation，co-citation analysis，literature coupling，and word cluster-based methods for identifying research fronts；at the same time，research front identification based on multi-source data，multi-dimensional indicators and machine learning algorithms is a new direction of the future research. On the basis of the above analysis，this paper summarizes the shortcomings of different types of research front identification methods and the existing universal problems，and looks forward to the future research priorities.

Keywords：research front expert interpretation citation analysis word cluster analysis multi-source data multi-dimensional indicators